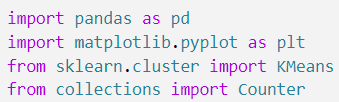
Ερώτημα 2

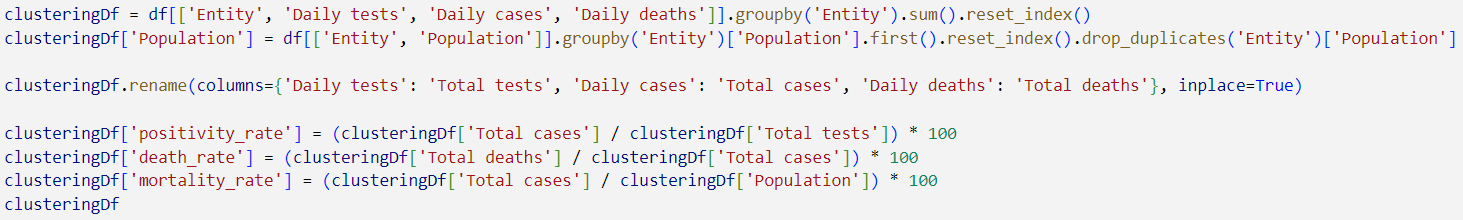
Αρχικά κάνουμε import τις κατάλληλες βιβλιοθήκες και κλάσεις που θα χρειαστούμε.



Την pandas για την επεξεργασία των .csv αρχείων ως dataframes, την matplotlib.pyplot για τη σχεδίαση των γραφικών παραστάσεων, την κλάση KMeans από τη βιβλιοθήκη sklearn.cluster για την συσταδοποίηση των δεδομένων μας και τέλος την κλάση Counter από τη βιβλιοθήκη collections για τη καταγραφή του αριθμού των entities που περιέχονται σε κάθε cluster.



Στη συνέχεια φορτώνουμε το dataset αφού έχει περάσει από την προ-επεξεργασία στο προηγούμενο ερώτημα.



Έπειτα δημιουργούμε ένα clustering dataframe όπου επιλέγουμε τις στήλες Entity, Daily tests, Daily cases και Daily deaths και κάνουμε sum aggregation με βάση το Entity και τέλος reset index.

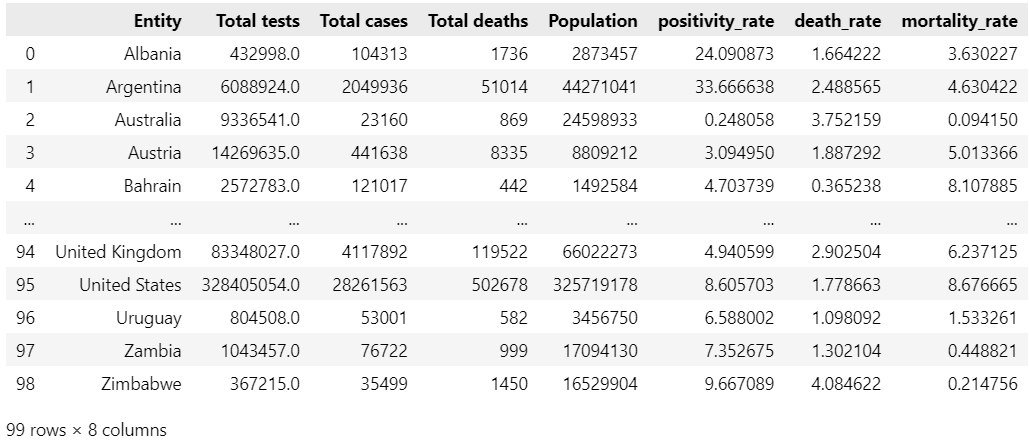
Δημιουργούμε μια νέα στήλη για το Population όπου παίρνουμε τον πληθυσμό για κάθε Entity και ακολουθούν οι κατάλληλες εντολές reset index και drop duplicates για να μετατραπεί το dataframe σε μία στήλη. Συνεχίζοντας αλλάζουμε το όνομα των στηλών και υπολογίζουμε τα απαραίτητα μεγέθη που ζητάει η εκφώνηση.

Το ποσοστό θετικότητας (positivity rate) ισούται με το πηλίκο των συνολικών cases διά των συνολικών tests επί 100%

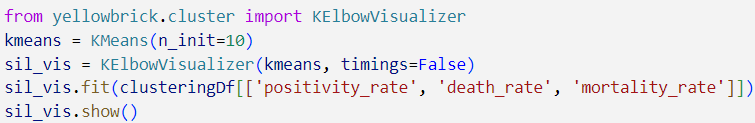
Το ποσοστό θνησιμότητας (death rate) ισούται με το πηλίκο των συνολικών θανάτων διά των συνολικών cases επί 100%

Ο αριθμός κρουσμάτων σε σχέση με τον πληθυσμό της χώρας (mortality rate) ισούται με το πηλίκο των συνολικών cases διά του πληθυσμού κάθε χώρας επί 100%

Το dataframe που προκύπτει είναι το εξής:

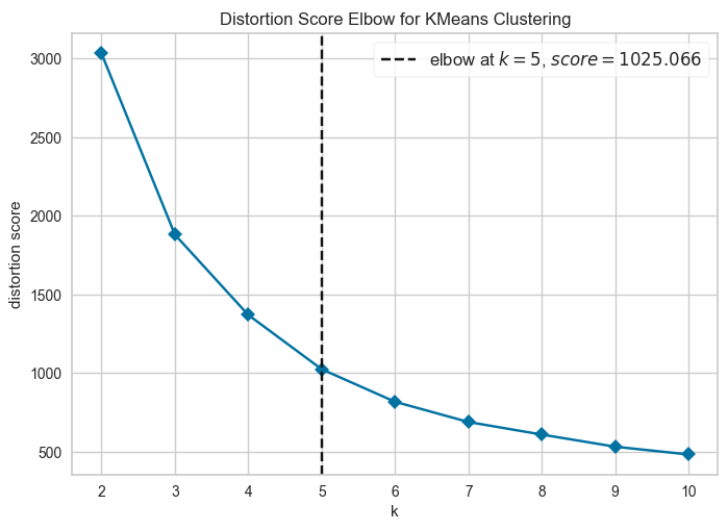


Ο αλγόριθμος συσταδοποίησης που χρησιμοποιήθηκε είναι ο KMeans καθώς τα δεδομένα εισόδου (positivity rate, death rate και mortality rate) είναι ευκλείδια. Δεν χρησιμοποιήθηκε το longitude και latitude καθώς είναι γνωστό ότι ακολουθούν σφαιρική γεωμετρία που είναι ένα παράδειγμα μη ευκλείδιας γεωμετρίας και ο KMeans δέχεται μόνο ευκλείδια δεδομένα.



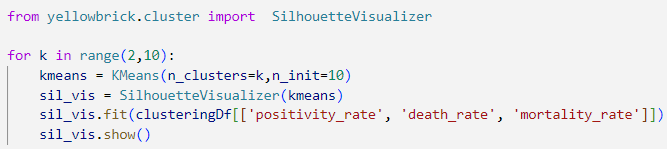
Ο KElbowVisualizer υλοποιεί την elbow μέθοδο για να μας βοηθήσει να επιλέξουμε το βέλτιστο αριθμό συστάδων προσαρμόζοντας το μοντέλο με ένα διάστημα τιμών για το K. Αν το διάγραμμα μοιάζει με ένα χέρι, τότε ο «αγκώνας» (το σημείο κάμψης της καμπύλης) είναι μια καλή ένδειξη ότι το υποκείμενο μοντέλο προσαρμόζεται καλύτερα σε αυτό το σημείο. Στο παρακάτω γράφημα ο «αγκώνας» θα σημειωθεί με μια διακεκομμένη γραμμή.

Από προεπιλογή, η μετρική της παραμέτρου βαθμολόγησης έχει οριστεί ως το distortion (παραμόρφωση), η οποία υπολογίζει το άθροισμα των τετραγωνικών αποστάσεων από κάθε σημείο προς το κέντρο που του έχει ανατεθεί.



Από το γράφημα βλέπουμε ότι ο KElbowVisualizer μας υπέδειξε ότι ο βέλτιστος αριθμός συστάδων είναι 5.

Ας εξετάσουμε και μια άλλη μετρική που ονομάζεται Silhouette Coefficient.



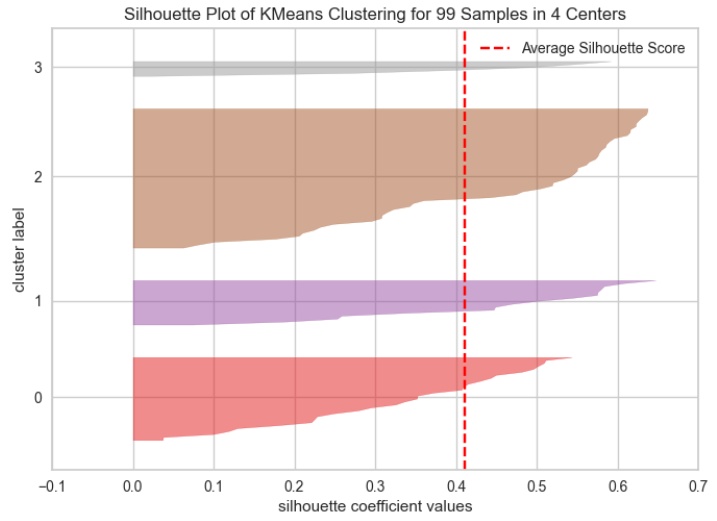
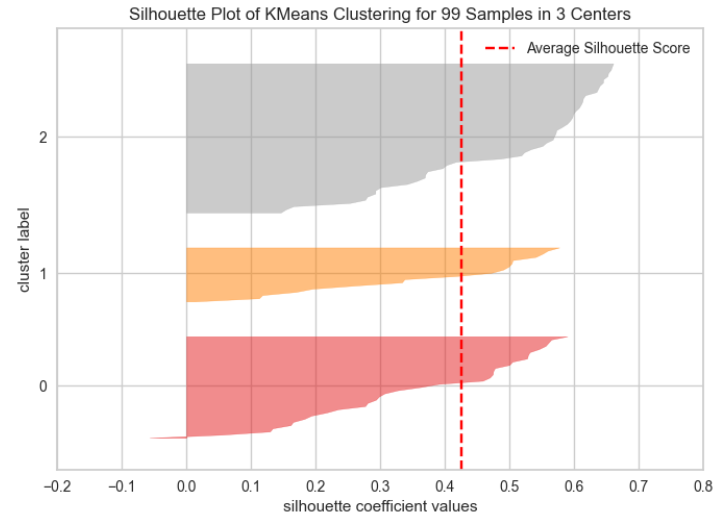
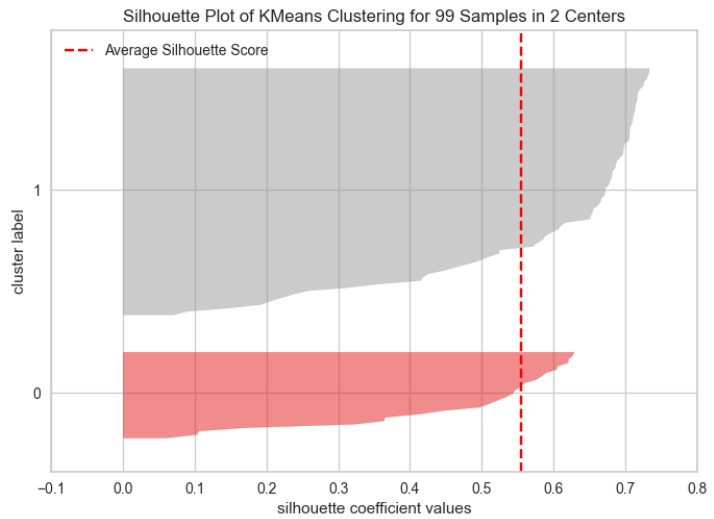
Χρησιμοποιούμε τώρα τον SilhouetteVisualizer και θέλουμε να παρατηρήσουμε από τα γραφήματα που θα προκύψουν για k από 2 έως 10 ποιος αριθμός συστάδων είναι ο βέλτιστος.

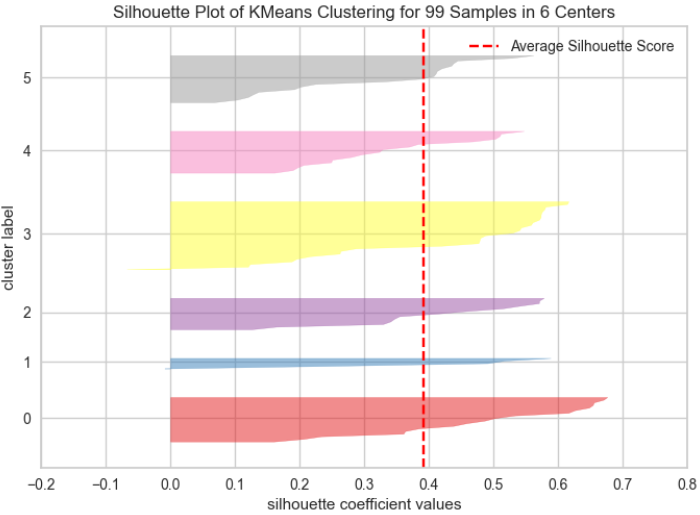
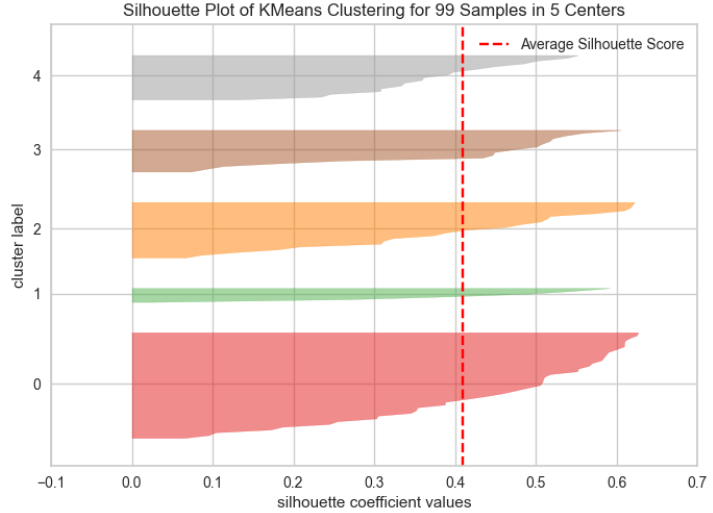
Ο SilhouetteVisualizer εμφανίζει το silhouette coefficient για κάθε δείγμα ανά συστάδα αξιολογώντας οπτικά την πυκνότητα και τον διαχωρισμό μεταξύ των συστάδων. Η βαθμολογία υπολογίζεται με τη μέση τιμή του silhouette coefficient για κάθε δείγμα, ο οποίος υπολογίζεται ως η διαφορά μεταξύ της μέσης απόστασης εντός συστάδα και της μέσης απόστασης πλησιέστερης συστάδας για κάθε δείγμα, κανονικοποιημένη από τη μέγιστη τιμή. Αυτό παράγει μια βαθμολογία μεταξύ -1 και +1, όπου βαθμολογίες κοντά στο +1 υποδηλώνουν υψηλό διαχωρισμό και βαθμολογίες κοντά στο -1 υποδηλώνουν ότι τα δείγματα μπορεί να έχουν ανατεθεί σε λάθος συστάδα.

Στα γραφήματα του SilhouetteVisualizer οι συστάδες με υψηλότερες βαθμολογίες έχουν πιο wide silhouettes, αλλά οι συστάδες που είναι λιγότερο συνεκτικές θα υπολείπονται της μέσης βαθμολογίας από όλες τις συστάδες, η οποία απεικονίζεται ως μια κατακόρυφη κόκκινη γραμμή.

Άρα για να διαλέξουμε το κατάλληλο K θέλουμε τα clusters να απεικονίζονται ομοιόμορφα wide, να είναι πάνω από την κόκκινη διακεκομμένη γραμμή και να μην πηγαίνουν πίσω από το 0 που δηλώνει ότι έχει γίνει λάθος στο clustering και κάποια δείγματα κανονικά ανήκουν σε άλλη συστάδα.

Οι γραφικές που προκύπτουν από το παραπάνω κομμάτι κώδικα είναι:



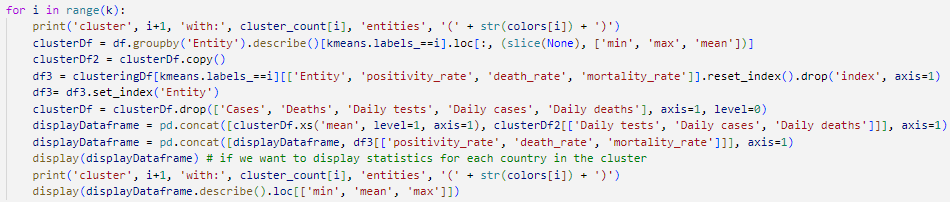


Παρατηρούμε ότι σε όλα τα διαγράμματα όλα τα clusters περνάνε την κόκκινη κατακόρυφη γραμμή. Στα διαγράμματα με K=3, 6 υπάρχουν clusters όπου έχει γίνει λάθος επειδή συνεχίζουν και πριν το 0 οπότε τα απορρίπτουμε. Στο διάγραμμα με K=2 υπάρχει μεγάλο imbalance αφού το ένα είναι πολύ πιο παχύ από το άλλο. Τα διαγράμματα με K=4 και K=5 είναι πολύ παρόμοια αλλά φαίνεται στο διάγραμμα με K=4 τα clusters να μην είναι τόσο ομοιόμορφα στο πάχος κι αφού και από την προηγούμενη μετρική με το KElbowVisualizer βρήκαμε ότι το βέλτιστο K=5 τότε αυτό θα επιλέξουμε τελικά.

Τέλος υλοποιούμε ένα 3d scatter plot για κάθε πιθανό K των σημείων κάθε συστάδας ώστε να δούμε καλύτερα τη διάταξη τους στο χώρο και να παρατηρήσουμε καλύτερα την ομαδοποίηση που έχει πραγματοποιηθεί.

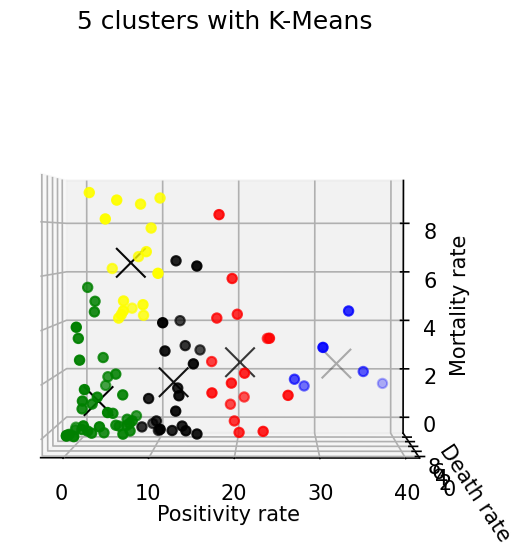
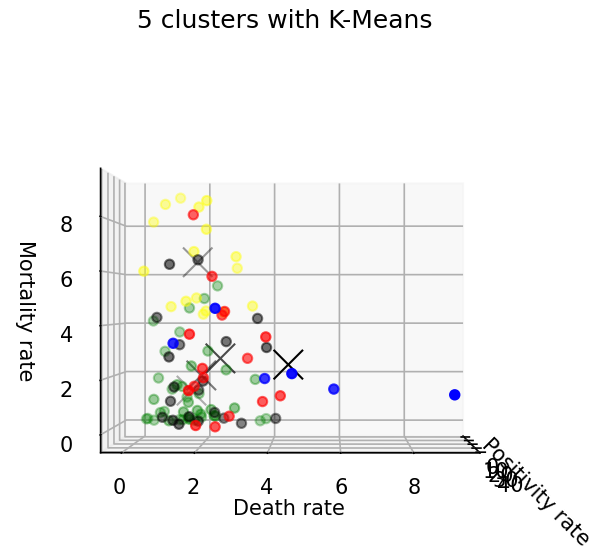


Κάνουμε πάλι clustering και φτιάχνουμε τα plots ορίζοντας κατάλληλα χρώματα για τα δείγματα κάθε συστάδας, τα απαραίτητα labels και το που βρίσκονται τα κεντροειδή.



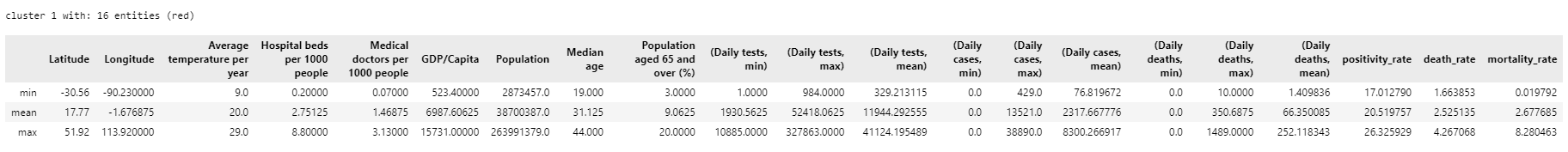
Για κάθε συστάδα βρίσκουμε τα απαραίτητα στατιστικά στοιχεία που χρειαζόμαστε και ποιες χώρες βρίσκονται σε αυτή.

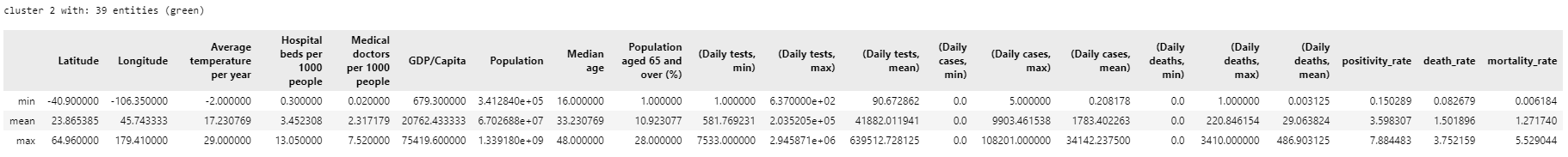
Εμείς θα επικεντρωθούμε στην γραφική για K=5 clusters που βρήκαμε ότι είναι το βέλτιστο

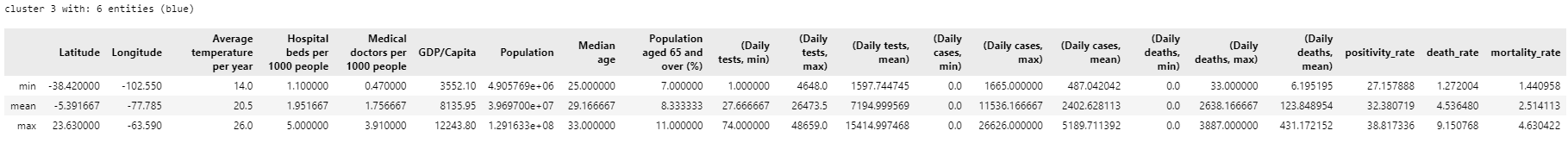
 

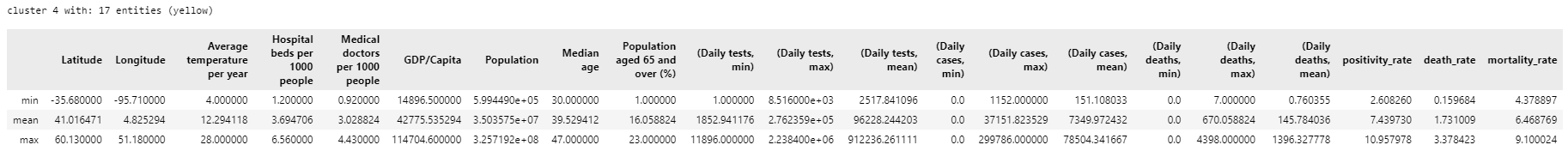
Από τα διαγράμματα φαίνεται ότι το μπλε cluster έχει υψηλές τιμές positivity rate και death rate ενώ το κίτρινο cluster φαίνεται ότι έχει υψηλές τιμές mortality rate.

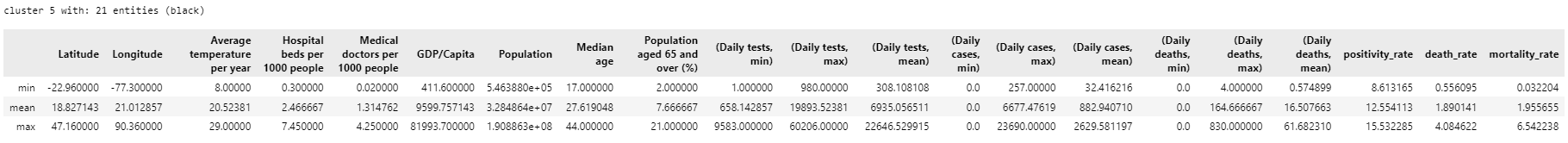
Οι στατιστικές τιμές για κάθε cluster παρατίθενται παρακάτω.





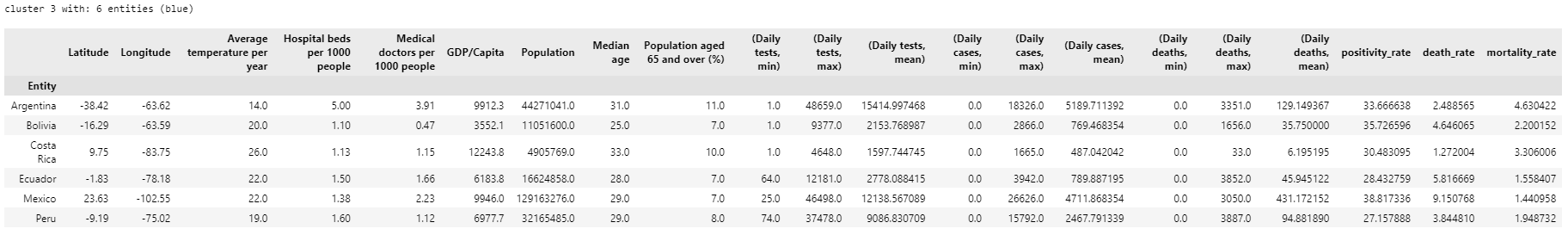






Θα απαντήσουμε πρώτα στην ερώτηση αν υπάρχουν χώρες που ξεχωρίζουν αρνητικά. Για να ξεχωρίσει μια συστάδα αρνητικά θα πρέπει να έχει υψηλές mean values και στα τρία πεδία (positivity rate, death rate και mortality rate).

Παρατηρούμε ότι η συστάδα 3 έχει τη μεγαλύτερη μέση τιμή για positivity rate και death rate από όλες τις άλλες συστάδες και την τρίτη μεγαλύτερη μέση τιμή για mortality rate άρα δίκαια μπορεί να χαρακτηριστεί ως η συστάδα που ξεχωρίζει αρνητικά. Άρα οι 6 χώρες που περιέχει ξεχωρίζουν αρνητικά και είναι οι εξής:

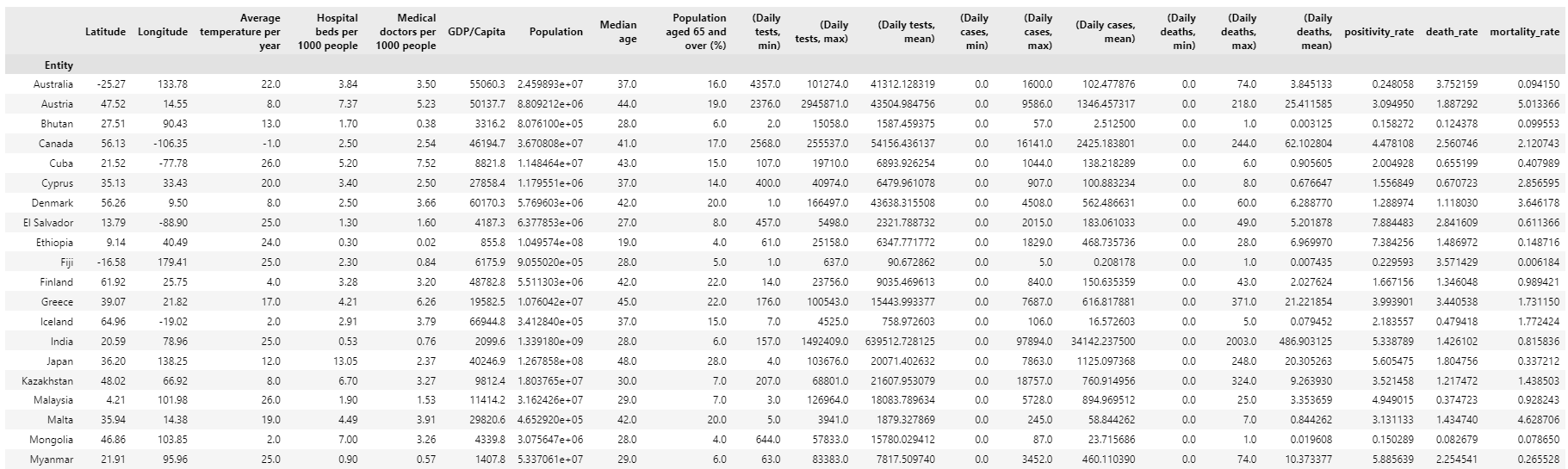
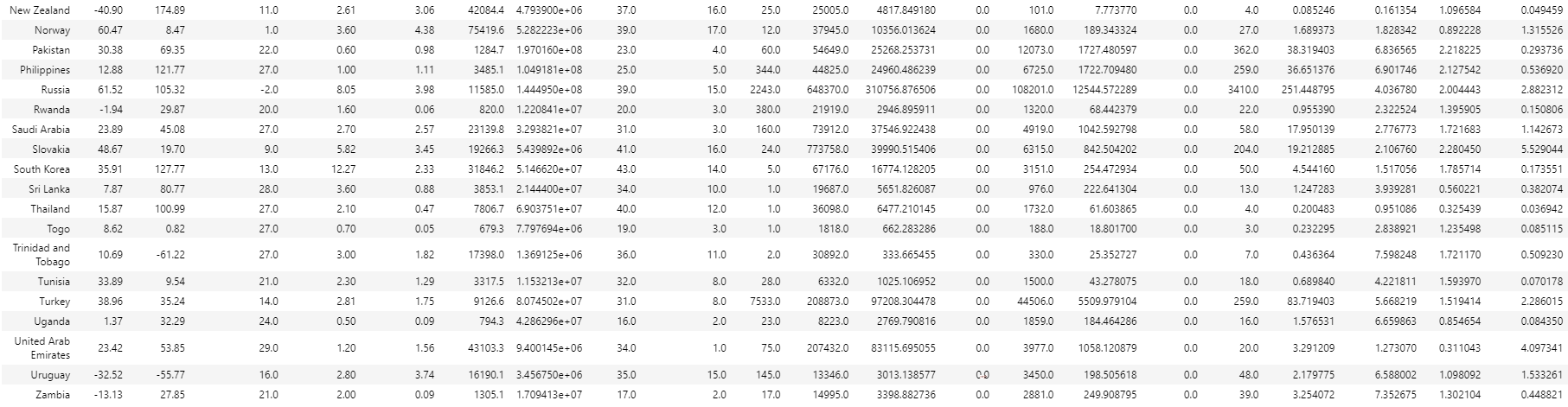


Τα κοινά χαρακτηριστικά που έχουν είναι ότι βρίσκονται όλες εκτός από το μεξικό στη νότια αμερική (το οποίο μεξικό βρίσκεται πολύ κοντά γεωγραφικά στην κόστα ρίκα). Έχουν παρόμοια θερμοκρασία και λίγο αριθμό κρεβατιών και γιατρών. Επιπλέον έχουν χαμηλό GDP/Capita.

Η χώρα με την χειρότερη κατάσταση που μπορεί να θεωρηθεί ως outlier είναι το Μεξικό καθώς συμβάλει στη μεγαλύτερη τιμή του positive και death rate για τη συστάδα του.

Τώρα θα απαντήσουμε στην ερώτηση αν υπάρχουν χώρες που ξεχωρίζουν θετικά. Για να ξεχωρίσει μια συστάδα θετικά θα πρέπει να έχει χαμηλές mean values και στα τρία πεδία (positivity rate, death rate και mortality rate).

Παρατηρούμε ότι η συστάδα 2 έχει τη μικρότερη μέση τιμή και για τα τρία πεδία άρα δίκαι μπορεί να χαρακτηριστεί ως η συστάδα που ξεχωρίζει θετικά. Άρα οι 39 χώρες που περιέχει ξεχωρίζουν θετικά και είναι οι εξής:

Επειδή είναι πολλές οι χώρες δεν μπορούμε να εντοπίσουμε κοινά χαρακτηριστικά μεταξύ τους. Αυτό που μπορούμε να κάνουμε είναι να συγκρίνουμε το μέσο όρο μερικών χαρακτηριστικών με τον αντίστοιχο μέσο όρο αυτών των χαρακτηριστικών για τη συστάδα 3.

Από τις παραπάνω εικόνες για τα στατιστικά των δύο χωρών μπορεί να γίνει αντιληπτό ότι η ιατρική περίθαλψη στις χώρες της συστάδας 2 είναι καλύτερη από αυτή των χωρών της συστάδας 3 αφού κατά μέσο όρο έχουν περισσότερους γιατρούς και κρεβάτια. Επίσης κατά μέσο όρο το GDP/Capita είναι 2 φορές μεγαλύτερο της συστάδας 2 από τη συστάδα 3.